Санкт-Петербургский
государственный университет
Кафедра теории управления

Журавлева Дарья Ильинична

Магистерская диссертация

Использование современных методов анализа данных в задачах биомеханики глаза

Направление 01.04.02
Прикладная математика и информатика
Магистерская программа **Прикладная математика и информатика в задачах медицинской диагностики**

Научный руководитель
к. ф-м. н., доцент
Воронкова Е. Б.

Санкт-Петербург
2018

Содержание

Введение

Глаукома является главной причиной необратимой слепоты во всем мире [1]. Открытоугольная глаукома (ОУГ) составляет более 90% случаев этого заболевания. Патофизиология глаукомы сложна и недостаточно изучена [2]. Согласно *механической* теории, повреждение ганглиозных клеток сетчатки возникает в результате повышения внутриглазного давления (ВГД); согласно же *сосудистой* теории, в патогенезе глаукомы важную роль играют иные факторы, а именно — пониженная перфузия глаза и, как следствие, ишемия диска зрительного нерва (ДЗН) и сетчатки [3]. В пользу *сосудистой* теории свидетельствует такие факты, как увеличение частоты кровоизлияний в диск зрительного нерва, мигреней и других осложнений [4] при глаукоме, значения ВГД в пределах нормы у более половины пациентов с глаукомой при первом посещении. Таким образом, появляется все больше доказательств того, что патогенез глаукомы связан с сосудистой дисфункцией [5].

Недавние исследования [5] показали, что кровоизлияние в диск зрительного нерва и перипапиллярная атрофия, — оба связаны с ускорением прогрессирования глаукомы. В клинической практике применяется широкий спектр устройств, чтобы детектировать и квантифицировать этот процесс: тест поля зрения (*visual field testing*), оптическая когерентная томография (*optical coherence tomography — OCT*) применяется для определения толщины слоя нервных волокон сетчатки в перипапиллярной области, Heidelberg retina tomography (HRT) измеряет стереометрические параметры и др. [2]. Даже при имеющихся в настоящее время диагностических и терапевтических возможностях многим пациентам так и не поставлен диагноз или заболевание продолжает прогрессировать, несмотря на лечение [6].

До недавнего времени отсутствовал воспроизводимый и надёжный *in vivo* метод количественной оценки кровоснабжения и микрососудистых сетей [7]. В конце 2016 года FDA (Food and Drug Administration, USA) одобрило оптическую когерентную томографию с ангиографией (*optical coherence tomography angiography — OCTA*) для клинической практики [8]. В октябре 2016 года Heidelberg Engineering выпустил OCT Angiography (OCTA) Module для своего основного инструмента — SPECTRALIS [9]. OCTA — это неинвазивный метод визуализации, который теперь позволяет одновременно визуализировать *in vivo* морфологию и сосудистую структуру глаза. В отличие от обычных структурных OCT изображений, OCTA изображения могут быть сделаны послойно, и, таким образом, можно исследовать кровоснабжение различных слоёв сетчатки. Поскольку первичная открытоугольная глаукома является оптической нейропатией с возможным сосудистым компонентом в ее многофакторной этиологии, OCTA изображения чрезвычайно интересны и полезны в диагностике данного заболевания [10].

Постановка задачи

Цели:

- Разработать метод обработки медицинских изображений, полученных на аппарате the SPECTRALIS® OCT2 platform с модулем OCT Angiography Module (Heidelberg Engineering GmbH), для количественной оценки микрокапиллярного кровотока.
- Оценить диагностическую информативность показателей, полученных с помощью анализируемых изображений.
- Оценить возможность применения полученной оценки кровенаполнения в клинической практике (при постановке диагноза).

Задачи:

- Создать алгоритм обработки изображений типа OCTAngio B-scan;
- Изучить полученные данные;
- Произвести отбор (или создание новых) признаков для решения задачи классификации;
- Решить задачу многоклассовой классификации на основе полученных данных.

Обзор литературы

ОСТА является новым методом диагностики. Результаты некоторых уже проведённых исследований позволяют рассматривать ОСТА как дополнительный инструмент для помощи врачам в клинической практике в обнаружении и контроле глаукомы [8]. В работе с ОСТА различают B-scan изображения и A-scan (или *en face*) изображения; схема представлена на рисунке 1. Как правило, в клинической практике врачи работают с A-scan

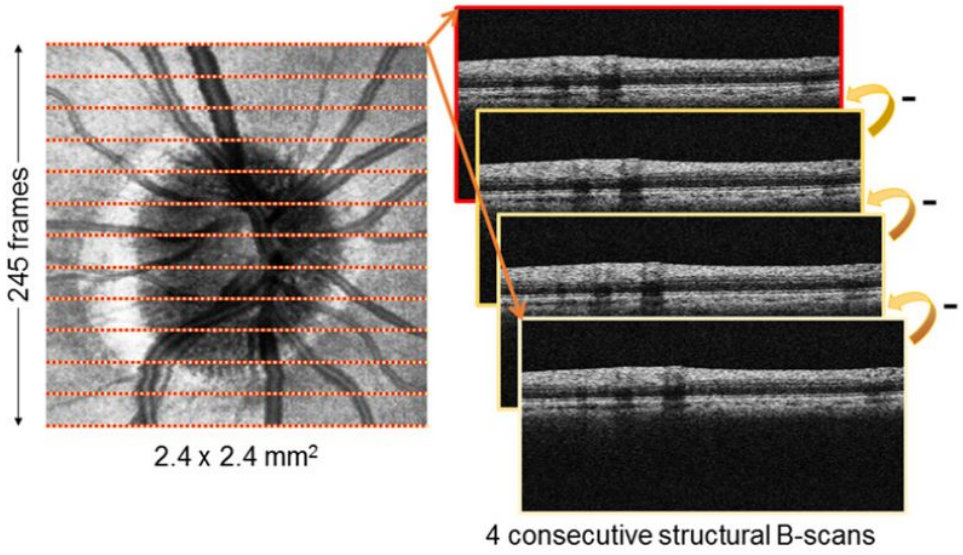

Рис. 1: Схема изображений ОСТА. Слева — A-scan, справа — изображения структурных B-scan. Источник: [4].

изображениями. Информация о некоторых исследованиях представлена в таблице на следующей странице. В основе ОСТА — улавливание сигнала движения частиц материала в сетчатке. В большинстве случаев — это красные кровяные тельца в сосудах сетчатки [8].

Плотность сосудов (*vessel density*) и индекс потока (*flow index*) — два наиболее часто используемых показателя [4, 6, 7, 12], которые могут быть получены по A-scan изображениям. Плотность сосудов определяется как доля в процентах площади, занимаемой кровеносными сосудами; индекс потока рассчитывается как средний сигнал потока в интересующей области. Индекс потока содержит информацию как о площади, занимаемой сосудом, так и о скорости кровотока. При проиллюстрированном на рисунке 1 подходе связь между потоковым сигналом и реальным потоком крови не идентифицируема из-за влияния накопления сигнала при высокой скорости кровотока [13]. В результате ОСТА изображения используются как статические карты ветвистости сосудов [8].

В данной работе предлагается алгоритм обработки B-scan изображения для оценки микрокапиллярного кровотока.

Год	Название статьи	Цель	Данные	Методы	Результат
2018	Correlation of flow density, as measured using optical coherence tomography angiography, with structural and functional parameters in glaucoma patients	сравнить кровоснабжение макулы и головки зрительного нерва, измеренное с помощью OCTA (en face), у пациентов с глаукомой и у здоровых людей	34 глаза тридцати четырёх пациентов с ОУГ и 35 глаз тридцати пяти здоровых пациентов	корреляция Спирмена с различными структурными и функциональными параметрами	плотность кровотока (en face) в группе глаукомы была значительно ниже по сравнению с контрольной группой. Наиболее сильная корреляция обнаружена с минимальной шириной неврального ободка (BMO-MRW).
2017	Vessel density analysis in patients with retinitis pigmentosa by means of optical coherence tomography angiography	описать сосудистые аномалии у пациентов, страдающих пигментным ретинитом (РП) средствами OCTA	шестнадцать пациентов (32 глаза)	исследуемые показатели: плотность поверхностного капиллярного сплетения (SCP), глубинного капиллярного сплетения (DCP) в области хориокапиллярного слоя (CC). плотность сосуда выражалась как отношение между пикселями сосуда и общей площадью. Наборы полученных значений сравнивались с контролями с помощью статистического анализа.	обнаружена статистически значимая разница в SCP и в DCP между пациентами и контролем; нет таковой в CC
2016	Optic disc perfusion in primary open angle and normal tension glaucoma eyes using optical coherence tomography-based microangiography	исследовать перфузию диска зрительного различия в норме, первичной ОУГ и глаукоме с нормальным давлением (ГНД) глаза с помощью optical microangiography (OMAG, vascular en face images) на основе OCTA метод	28 норма, 30 первичная ОУГ, 31 ГНД	для количественной оценки перфузии диска зрительного нерва создано три количественных измерения: поток, плотность площади сосудов и нормализованный поток; проведён одномерный регрессионный анализ между показателями кровотока и функциональными и структурными измерениями	по сравнению с нормальными глазами, глаза первичной ОУГ и глаза ГНД имели значительно более низкий поток, низкую плотность площади сосудов и низкий нормализованный поток в предламинарной области ДЗН.
2017	Peripapillary perfused capillary density in primary open-angle glaucoma across disease stage: an optical coherence tomography angiography study	оценить перипапиллярную перфузионную плотность капилляров (PCD) при первичной ОУГ	60 глаз различной стадии первичной ОУГ и 24 глаз контроля	PCD рассчитывался в процентах как отношение количества пикселей кровенаполненных капиляров, к общему количеству пикселей в соответствующей области интереса (ROI); анализ ковариаций использовался для сравнения PCD среди групп и контроля	PCD постепенно уменьшался по мере прогрессирования первичной ОУГ на всех ROI
2016	Quantifying microvascular density and morphology in diabetic retinopathy using spectral-domain optical coherence tomography angiography	количественно оценить изменения микроциркуляторного русла сетчатки при диабетической ретинопатии (ДР) с помощью SD-OCTA	84 глаза с ДР и 14 здоровых	для расчета показателей микрососудистой плотности и морфологии использовалась полуавтоматическая программа; статистический анализ проводился с использованием t-критерия Стьюдента или дисперсионного анализа	сосудистые изменения при ДР могут быть объективно и достоверно охарактеризованы SD (плотность скелета), VD (плотность сосуда), FD (морфология сосуда) и VDI (индекс диаметра сосуда). В целом, снижение плотности капилляров (SD и VD), сложность ветвления (FD) и увеличение среднего сосудистого диаметра (VDI) были связаны с ухудшением течения заболевания.
2016	Optical coherence tomography angiography vessel density in healthy, glaucoma suspect, and glaucoma eyes	сравнить толщину слоя нервных волокон сетчатки (RNFL) и измерения сосудистой сети сетчатки по OCTA у здоровых пациентов, с подозрением на глаукому и у больных глаукомой	261 глаз всего	AUROC (area under receiver operating curve) использовался для оценки точности диагностики	измерения OCTA отражают изменения, связанные с патофизиологией ОУГ

Теория классификации

Задача классификации состоит в построении правила (классификатора), которое наилучшим образом разделяет объекты на заранее известные классы, при этом совершая как можно меньше ошибок. Предполагается, что ответ известен для каждого объекта.

По количеству классов, к которым могут принадлежать объекты, разделяют задачу **бинарной** классификации и задачу **многоклассовой** классификации.

В данной главе будем полагать, что нам известны объекты и ответы на них. Введем следующие обозначения:

- N — количество объектов; M — количество признаков; S — количество классов,
- x — объект, математически представленный вектором, в котором каждая компонента — значение признака, всего компонент M,
- $X_{[N \times M]}$ — матрица объектов, где каждая строка — объект, а каждый столбец — признак,
- y — ответ на объекте x, математически представленный целым числом, $y \in \mathbb{Z}$,
- Y — вектор-столбец ответов на объектах X,
- $\hat{y}, \hat{Y}$ — предсказания, полученные неким алгоритмом классификации, для одного и всех объектов, соответственно.

Метрики качества классификации

При решении задачи классификации важно оценить, насколько близки предсказания, полученные выбранным алгоритмом классификации, к правильным ответам, т.е. как близок вектор предсказаний $\hat{Y}$ к вектору правильных ответов Y.

Бинарная классификация. Обозначим классы, на которые необходимо разделить множество объектов, как 1 и -1, и введем общепринятые обозначения: $\boldsymbol{TP}$ — количество истинно положительных ответов, $\boldsymbol{FP}$ — количество ложно положительных ответов, $\boldsymbol{FN}$ — количество ложно отрицательных ответов, $\boldsymbol{TN}$ — количество истинно отрицательных ответов [14].

Для оценки качества классификатора могут быть использованы следующие метрики [14]:

- Совпадение

$$Accuracy(Y, \hat{Y}) = \frac{1}{N} \sum_{i=1}^{N} 1(y_i == \hat{y}_i), \tag{1}$$

где $1(\cdot)$ - индикаторная функция.

Формула (1) вычисляет долю верных предсказаний. Оценки данной метрики могут вводить в заблуждение при несбалансированности классов, т.е. когда объектов одного класса значительно больше, чем другого. В таком случае разумно установить границу в области значений метрики, равную доле самого многочисленного класса: если метрика предсказаний алгоритма ниже этой границы, то алгоритм некорректен.

- Точность

$$Precision(Y, \hat{Y}) = \frac{TP}{TP + FP} \tag{2}$$

- Полнота

$$Recall(Y, \hat{Y}) = \frac{TP}{TP + FN} \tag{3}$$

Метрика (2) выделит те алгоритмы, которые в своих предсказаниях меньше ошибаются на объектах класса -1, в то время как (3) выделит те алгоритмы, которые в своих предсказаниях меньше ошибаются на объектах класса 1.

- F-мера

$$F_\beta(Y, \hat{Y}) = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{4}$$

Формула (4) представляет собой гармоническое среднее между (2) и (3). При $\beta = 1$ точность и полнота имеют одинаковый вес.

		Y	
		1	-1
$\hat{Y}$	1	TP	FP
	-1	FN	TN

Таблица 1: Поясняющая таблица к формулам (2) и (3).

Многоклассовая классификация. Когда классов не два, а больше, встаёт вопрос о том, какие метрики придумать или как распространить метрики бинарной классификации на задачи многоклассовой классификации. Ниже перечислены метрики, которые обычно используют при разделении объектов на три и более класса [15].

- Сумма квадратов отклонений

$$MSE(Y, \hat{Y}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2. \tag{5}$$

Оценки метрики (5), также как и оценки метрики (1), могут вводить в заблуждение при несбалансированности классов.

- Бинарная классификация каждого класса
Для каждого класса выполняется процесс бинарной классификации с использованием любой из метрик (1)-(4). Затем полученные S оценок усредняют, применяя одну из стратегий [15]:

1. **Макро** — вычисляет среднее значение оценок метрик бинарных классификаций, придавая одинаковый вес каждому классу;
2. **Взвешенная** — вычисляет взвешенное среднее; при равенстве всех весов 1 результат тот же, что и при макро стратегии;
3. **Микро** — позволяет каждой задаче бинарной классификации внести вклад в общую оценку. Например, для трёх классов усредненная оценка может быть вычислена по формуле

$$Micro_{Precision_3} = \frac{TP_1 + TP_2 + TP_3}{(TP_1 + TP_2 + TP_3) + (FP_1 + FP_2 + FP_3)}.$$

Эту метрику желательно использовать в задачах, когда необходимо выделить алгоритмы лучше справляющиеся с классификацией не самого многочисленного класса.

Выбор признаков

Когда признаков у объектов больше трёх ($M > 3$), необходимо решить два связанных вопроса

1. Как получить признаки, имеющие наибольшую описательную способность? Возможно, эти признаки нужно выбрать из доступных, или составить комбинации из доступных.

2. Как отображать объекты? Этот вопрос является частным случаем первого, так как необходимо искать три (в случае пространства) или два (в случае плоскости) таких признака.

Опишем кратко основные методы, позволяющие получить ответы на эти вопросы.

Фильтрующие методы. Эта группа методов ранжирует признаки по некоторому критерию, после чего пользователь может выбрать несколько первых в полученном списке. Возможными критериями для ранжирования являются [16]:

- Критерий Пирсона

$$R(feature) = \frac{cov(X_{feature}, Y)}{sqrtvar(X_{feature}) \cdot var(Y)} \tag{6}$$

  может обнаруживать только линейные зависимости между признаком и ответом.

  Здесь и далее, $feature$ — признак, $X_{feature}$ — столбец матрицы объектов X, соответствующий данному признаку, Y — вектор-столбец ответов на объектах X.

- Взаимная информация, определяемая как

$$I(feature) = H(Y) - H(Y|X_{feature}) \tag{7}$$

  для дискретных $X_{feature}$ и Y через функцию энтропии $H(\cdot)$ [17].

  Для непрерывных $X_{feature}$ и Y взаимная информация вычисляется по формуле

$$I(feature) = \int f_Y(t) log\left(\frac{f_Y(t)}{g_{X_{feature}(t)}}\right) \tag{8}$$

  где f и g — плотности вероятности Y и $X_{feature}$, соответственно.

- MSE (Mean Squared Error)

$$MSE(feature) = \frac{1}{N}\sum_{i=1}^{N}(x_{i\ feature}^2 - y_i^2). \tag{9}$$

  Этот критерий требует нормализации как признаков, так и ответов.

- RELIEF

$$RELIEF(feature) = \sum_{i=1}^{N}(x_{i\ feature} - x_{nearest\ feature})^2 + \\ + \sum_{i=1}^{N}(x_{i\ feature} - x_{missnearest\ feature})^2. \tag{10}$$

  Критерий (10), так же, как и (9), требует предварительной нормализации признаков. Здесь $nearest$ — ближайший к объекту i объект того же класса, $missnearest$ — ближайший к объекту i объект другого класса [18].

Преимущество этих методов в том, что они не требуют больших вычислительных затрат и их легко применять. Главный недостаток состоит в том, что итоговое подмножество признаков может содержать избыточные признаки или, наоборот, быть слишком узким, так как выше перечисленные критерии (6)-(10) не учитывают корреляции признаков между собой [17].

Оптимизационные методы. Эти методы, фактически, оптимизируют целевую функцию, в нашем случае это оценка метрики классификации, перебирая некоторым алгоритмом комбинации признаков. Такие алгоритмы можно разделить на две группы:

- Алгоритмы последовательного поиска

  SFS (Sequential Feature Selection) алгоритм начинается с пустого набора признаков. Далее на каждой итерации добавляется тот признак, который даёт наилучшее значение целевой функции. Признаки добавляются до тех пор, пока не будет набрано требуемое количество.

  SBS (Sequential Backward Selection) — зеркальный алгоритм относительно SFS: начинается с набора, состоящего из всех признаков, и на каждой итерации исключается тот признак, удаление которого приводит к наилучшему значению функции. Алгоритм останавливается, когда набор состоит из требуемого количества признаков.

  SFFS (Sequential Floating Forward Selection) является гибкой версией SFS: перед включением признака, совершается шаг SBS.

  Plus-L-Minus-R алгоритм на каждой итерации позволяет включать до L и исключать до R переменных.

  Главным недостатком этой группы алгоритмов является зацикливание на вложенных подмножествах: в отобранных признаках могут оказаться сильно коррелирующие между собой.

- Эволюционные алгоритмы

  GA (Genetic Algorithm) — это семейство алгоритмов, которые получили такое название, потому что проводят поиск оптимального набора признаков, имитируя естественную эволюцию живых организмов. В основе их принципа действия лежит тот факт, что живые организмы эффективно адаптируются к изменяющимся условиям. Каждая итерация любого GA состоит из трёх этапов: отбор, кроссинговер и мутация. На этапе отбора наиболее приспособленные особи (имеющие лучшее значение целевой функции) получают больше шансов быть отобранными для размножения. Во время кроссинговера происходит обмен частями родительских решений в надежде получить более адаптированные решения. Мутация происходит путем случайного изменения одного или нескольких компонентов выбранного индивидуума. Мутации осуществляются для того, чтобы сохранить некото-

рое разнообразие в решениях.

Основным недостатком оптимизационных методов является количество вычислений, необходимое для получения подмножества объектов. На каждом этапе изменения количества отобранных признаков обучается классификатор и вычисляется оценка по выбранной метрике. Если количество выборок велико, то большая часть работы алгоритма тратится на обучение классификатора. В некоторых алгоритмах, таких как GA, одно и то же подмножество объектов может оцениваться несколько раз, так как точность классификатора для оцениваемых подмножеств не сохраняется. Еще один недостаток использования оценки метрики классификатора в качестве целевой функции состоит в том, что классификаторы склонны к переобучению, то есть в итоге результатом может стать хорошее значение целевой функции, но плохое подмножество признаков, обладающей плохой обобщательной способностью [17].

Встроенные методы. К ним относят алгоритмы встроенные непосредственно в сам классификатор. Примером такого алгоритма является линейный классификатор с регуляризацией lasso, подробно описанный в следующем разделе.

Трансформационные методы создают новые признаки на основе доступных [19]. Можно выделить следующие методы:

- PCA (Principal Component Analysis)
  Метод PCA используется для разложения многомерного набора данных в набор последовательных ортогональных компонент, которые объясняют наибольшую часть дисперсии. Главными компонентами являются первые собственные векторы выборочной ковариационной матрицы признаков (11), отсортированные по убыванию собственных чисел, которым они принадлежат.

$$W_{PCA} = C = cov(X, X), \tag{11}$$
$$c_{i,j} = \frac{1}{N}\sum_{i=1}^{N}((x_i - \bar{x}_i) \times (x_j - \bar{x}_j)),$$

  $\bar{\cdot}$ обозначает выборочное среднее.

  Из главных компонент составляется проекционная матрица; для получения образа объекта в этом пространстве, необходимо умножить его слева на эту матрицу. Перед применением этого алгоритма необходимо нормировать признаки.

- LDA (Linear Discriminant Analysis)
  Метод LDA пытается определить признаки, которые учитывают наибольшую дисперсию между классами, при этом признаки должны

быть непрерывными величинами. Проекционная матрица в данном случае составляется из собственных векторов матрицы $S_W^{-1}S_B$ (12).

$$W_{LDA} = S^{-1}_{within\ class} S_{between\ class}, \tag{12}$$

$$S_{within\ class} = S_W = \sum_{j=1}^{S}\sum_{i=1}^{N_j}(x_i^j - \mu_j)(x_i^j - \mu_j)^T,$$

N_j — количество объектов класса j,

$$S_{between\ class} = S_B = \sum_{j=1}^{S}(\mu_j - \mu)(\mu_j - \mu)^T,$$

μ_j — выборочное среднее в классе j,

μ — выборочное среднее по всем классам.

S_W отображает разброс значений признаков внутри каждого класса, S_B отображает разброс значений признаков между классами. Размерность полученного пространства всегда будет равна $(S-1)$. Как правило, для того, чтобы $S_{within\ class}$ не оказалась вырожденной, LDA применяют к матрице признаков после PCA.

- RICA (Reconstruction Independent Component Analysis)
  Данный метод, как и его оригинал — ICA, требует, чтобы данные были предварительно отшлифованы (*англ. whitened*). В классическом ICA решается следующая задача:

$$\min_W \sum_{i=1}^{S}\sum_{j=1}^{K} g(W_j x_i^T), \quad WW^T = 1,$$

где $W_{K\times S}$ — матрица весов, g — нелинейная выпуклая функция, W_j — строчка матрицы W, x_i - объект (строка) матрицы X, K — желаемое количество признаков в новом пространстве. Ортонормированное ограничение в ICA приводит к затруднениям в решении этой оптимизационной задачи. Метод RICA предлагает альтернативную постановку:

$$\min_W \frac{\lambda}{S}\sum_{i=1}^{S}||W^T W x_i^T - x_i^T||_2^2 + \sum_{i=1}^{S}\sum_{j=1} K g(W_j x_i^T),$$

где λ — сглаживающий коэффициент. Как правило, $g = \frac{1}{2}log(cosh(\cdot))$ [20].

Классификаторы

В данном разделе речь пойдёт о правилах (*классификаторах*, $a(x)$), которые могут быть использованы исследователем для того, чтобы добить-

ся наилучшего разделения объектов на заранее известные классы. Будем рассматривать классификаторы в контексте бинарной классификации, то есть будем полагать, что объекты разделены на два класса: -1 и 1.

Линейные

$$a(x) = sign(w_0 + xw^T), \tag{13}$$

где w_0, w_j — некоторые константы (веса). Геометрический смысл такого классификатора — это плоскость в пространстве признаков. Объекты, находящиеся по разные стороны от неё, принадлежат разным классам. Для нахождения w_0 и w_j, $j = 1, \ldots, S$, как правило, поступают так: полагают, что $Y \in R$, решают задачу регрессии, применяют правило (13). Причём при решении задачи регрессии используют вариации метрики MSE:

- Без регуляризации

$$MSE_{Classic}(Y, \hat{Y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i), \tag{14}$$

- Ridge регуляризация

$$MSE_{Ridge}(Y, \hat{Y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i) + \sum_{j=1}^{S} w_j^2, \tag{15}$$

- Lasso регуляризация

$$MSE_{Lasso}(Y, \hat{Y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i) + \sum_{j=1}^{S} |w_j|. \tag{16}$$

При решении задач классификации можно обнаружить такое явление, как переобучение алгоритма, то есть когда алгоритм не понимает закономерность получения ответа на объекте, а просто запоминает ответы. В линейных моделях такое явление легко определить: оно выражается в больших абсолютных значениях весов. Для борьбы с ним к классической метрике (14) добавляют норму, домноженную на коэффициент регуляризации: либо $||\cdot||_2$ норму (15), либо $||\cdot||_1$ норму (16). У метрики с lasso регуляризацией, помимо борьбы с переобучением, есть ещё одно важное свойство: она выделяет решения с меньшими значениями весов, имея при этом склонность обнулять веса, таким образом уменьшая число признаков [21]. Это пример встроенного метода выбора признаков.

SVM

Геометрический смысл support vector machine (SVM) — это гиперплоскость

или набор гиперплоскостей в пространстве. Наиболее хорошо разделяющая два класса гиперплоскость — та, расстояние от которой до ближайших объектов любого из классов максимально. Правило имеет тот же вид (13), что и для линейных классификаторов, но задача нахождения весов ставится иным образом:

$$\underset{w_j, w_0}{argmin} ||w||^2,$$
$$y_i(w_0 + x_i w^T) \geqslant 1.$$

SVM эффективен в случаях, когда классы действительно разделимы и когда объектов меньше, чем признаков; может быть некорректен в случаях, когда данные сильно зашумлены [22].

Бинарное дерево решений

Бинарное дерево решений — это правило, состоящее из последовательности простых решений (*листьев*) типа $[x^i < t]$, где x^i — i-ый признак объекта x, t — некоторое пороговое значение. В отличие от линейных классификаторов, дерево решений легче интерпретируется, но и легче переобучается. Дерево строится *жадным* способом — от корня к листьям: сначала выборка разбивается на две подвыборки, далее каждая из подвыборок разбивается на две и так далее. Выбор признака i, по которому проводится разбиение, выбирается путём минимизации *критерия ошибки*:

$$Q(X_M, i, t) = \frac{|X_L|}{|X_M|} H(X_L) + \frac{X_R}{X_M} H(X_R),$$

где X_M — разбиваемая выборка, X_L, X_R — выборки, которые получаются в результате разбиения, $|\cdot|$ — мощность множества, $H(\cdot)$ — критерий качества подмножества (отражает то, насколько высока вариация ответов в $[\cdot]$). В случае задачи классификации критерием качества может быть:

- критерий Джинни

$$H(X_M) = \sum_{m=1}^{M} p_m(1 - p_m) \tag{17}$$

- энтропийный критерий

$$H(X_M) = -\sum_{m=1}^{M} p_m ln(p_m), \text{ принимается } 0ln0 = 0. \tag{18}$$

- критерий ошибочной классификации

$$H(X_M) = 1 - \max_m p_m \tag{19}$$

В формулах (17), (18), (19) применено следующее обозначение

$$p_m = \frac{1}{|X_M|} \sum_{i \in X_M} [y_i = m]$$

Для всех трёх критериев верно, что они достигают своего минимального значения только когда все объекты в X_M принадлежат одному классу [23].

Как было отмечено ранее, деревья склонны сильно переобучаться. С этим явлением существует два метода борьбы:

1. Критерий останова. При очередном разбиении некоторой выборки на подвыборки каждая из подвыборок может быть либо листом (то есть должна быть далее разделена), либо вершиной (и здесь процесс построения заканчивается). Ограничения, выполнение которых приводит к окончанию процесса, могут быть на *количество объектов* (если количество объектов в подвыборке меньше некоторого уровня, то это — вершина) или на *глубину* (если выборка получена в результате R-ого последовательного принятия решений, то это — вершина; при этом R определяется заранее).

2. Стрижка деревьев. При таком подходе строится дерево максимальной глубины (т. е. до одного объекта в вершине), далее вершины объединяются, до тех пор, пока не достигнута будет желаемая глубина или пока в очередной вершине не будет достаточное количество объектов.

Композиция деревьев

Деревья решений способны выявлять сложные закономерности — в этом их несомненное преимущество, но их долго строить, они склонны к переобучению и могут сильно меняться при небольшом изменении выборки (неустойчивы к выбросам). Может ли композиция деревьев (то есть учёт ответов нескольких недообученных деревьев) улучшить классификацию?

- Bagging (Bootstrap AGGregation) подход
  Обратимся к следующей формуле:

  $$\text{Ошибка на новых объектах} = \text{Шум} + \text{Смещение} + \text{Разброс}. \quad (20)$$

  В (20) шум равен ошибке идеального алгоритма, смещение характеризует отклонение усреднённого по композиции ответа от ответа идеального алгоритма, разброс характеризует дисперсию ответов относительно ответа идеального алгоритма. Усреднение ответов недообученных деревьев уменьшает разброс и не меняет смещения, но чтобы это обеспечить, нам необходимы независимые алгоритмы. Примером реализации такой независимости является алгоритм *Random Forest* (случайный лес). В случайных лесах корреляция между деревьями снижается двумя путями:

1. признак, по которому производится разбиение выборки, выбирается не из всех возможных признаков, а лишь из их случайного подмножества размера K. Для задачи классификации рекомендуется $K = sqrt(S)$;
2. непосредственно bagging подход, то есть случайный выбор объектов, на которых обучается дерево композиции.

Преимущества случайного леса в том, что не происходит переобучения при росте числа базовых алгоритмов (деревьев решений), он хорошо параллелится и это удобно в практической реализации, так как для построения подвыборок используется бутстрап, то получается, что для любого объекта в композиции деревьев есть примерно 37% деревьев, которые этот объект не знают, таким образом, нет необходимости в кросс-валидации или отложенной выборки [24]. Основным недостатком является то, что составляющие композиции — глубокие деревья, причём каждое следующее дерево не зависит от качества предыдущего, таким образом, их нужно много для лучшей классификации.

- Boosting подход
  Другой подход состоит в следующем: каждый следующий алгоритм корректирует ответ предыдущего. Пример реализации такого подхода — AdaBoost (Adaptive Boosting). Идея состоит в том, что каждый следующий алгоритм обучается на тех объектах, на которых предыдущий ошибся. Таким образом, композиция из недообученных деревьев становится хорошим классификатором. Недостаток — сильная чувствительность к выбросам [25].

Результаты

Данные

В ФГБУ «Московский НИИ глазных болезней им. Гельмгольца» Минздрава России были проведены обследования пациентов, включающие тонометрию (измерение ВГД), пахиметрию роговой оболочки глаза (измерение толщины роговицы), исследование на анализаторе глазного ответа (ORA, Ocular Response Analyzer, США), а также оптическую когерентную томографию с функцией ангиографии (OCTA или ОКТ-А).

Для анализа были предоставлены следующие показатели:

- Параметры, измеренные на ORA

Pg ВГД по Гольдману (Pg, мм рт. ст.)

Po роговично-компенсированное ВГД (Po, мм рт. ст.)

Две биомеханические характеристики, отражающие вязкоэластические свойства ткани роговицы – корнеальный гистерезис и фактор резистентности роговицы

ФРР Фактор резистентности роговицы (ФРР, мм рт.ст.)

КГ Корнеальный гистерезис (КГ, мм рт.ст.)

ЦТР Центральная толщина роговица (ЦТР, мкм), измеренная на ORA [26]

Параметры Po, ФРР и КГ рассчитываются специальным алгоритмом, учитывающим вязко-эластические свойства роговицы.

- Параметры, измеренные по OCTA с помощью встроенных алгоритмов сегментации изображений:

ТРП Толщина решетчатой пластинки, (ТРП, мкм)

ГРП Глубина решетчатой пластинки, (ГРП, мкм)

SVL Толщина поверхностного сосудистого слоя сетчатки (surface vascular layer, SVL, мкм)

DVL Толщина глубокого сосудистого слоя сетчатки (deep vascular layer, DVL, мкм)

Другие параметры

Efrid Коэффициентом ригидности Фриденвальда, расчитанный по [27].

Диагноз Диагноз, поставленный пациенту врачом-офтальмологом.

Обработка изображений

Все изображения, представленные для обработки были получены в ФГБУ «Московский НИИ глазных болезней им. Гельмгольца» на аппарате SPECTRALIS® OCT2 platform OCT Angiography Module, Heidelberg Engineering GmbH (SPECTRALIS). Данный аппарат выполняет A-сканирования со скоростью 85 кГц с осевым разрешением 7 мкм и боковым разрешением 14 мкм. В аппарате используется алгоритм абсолютной деккорелляции полного спектра (FS-ADA). SPECTRALIS позволяет получать OCTA сканы (ОКТ ангиограммы) размером 3×3 мм. Стандартный OCTA скан состоит из 512 B-сканов, каждый из которых включает в себя 512 A-сканов. Дистанция между сканами — 6 мкм.

В процессе исследования анализировался один скан сагиттального среза (B-скан) диска зрительного нерва (ДЗН) и парапапиллярной зоны размером 3 мм. Кровенаполнение глазного яблока является динамическим процессом, поэтому для каждого пациента выбирался наиболее "яркий" из пяти сканов в зоне 1/2 ДЗН (то есть площадью выбора являлась зона диаметром 30 мкм), наиболее полно характеризующий наполнение кровью этой зоны в момент выполнения сканирования. Отбор такого скана производился врачом-офтальмологом.

SPECTRALIS производит автоматическую сегментацию слоев сетчатки глаза. На предоставленных изображениях были выделены верхний сплетениевидный слой (OPL) и мембрана Бруха (BM). Пример исходного изображения с выделенными слоями и схематичное строение сетчатки глаза представлены на рис. 2.

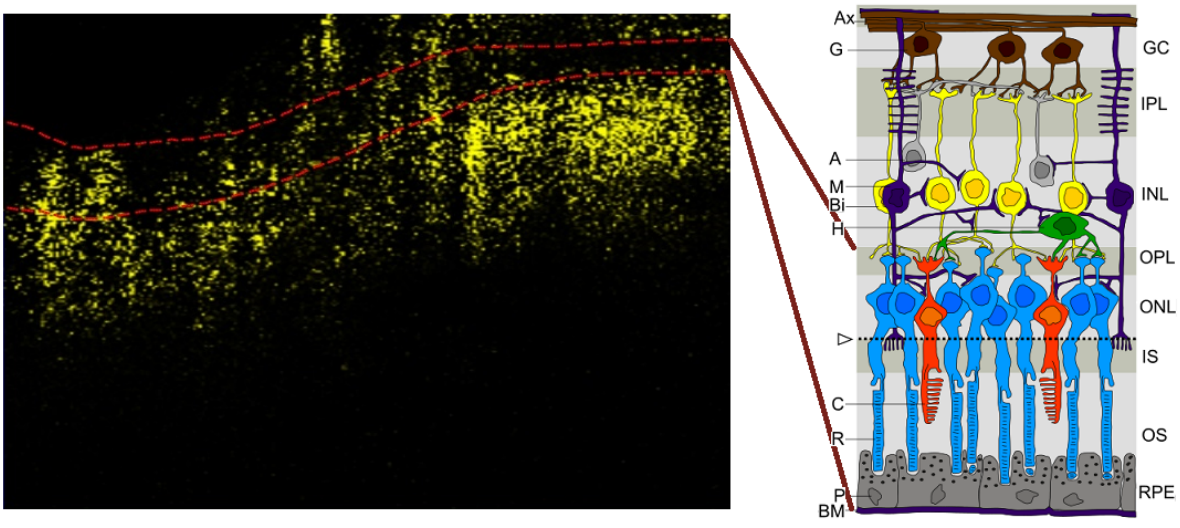

Рис. 2: Слева: Исходное изображение (OCTA B-скан) с выделенным кровенаполнением сосудов (желтый цвет). Верхняя красная линия — слой OPL (*верхний сплетениевидный слой*), нижняя красная линия — слой BM (*мембрана Бруха*). Справа: Слои сетчатки глаза (адаптировано с https://en.wikipedia.org/wiki/Retina).

Алгоритм обработки OCTA B-скан

Каждое изображение представляет собой трёхмерную матрицу или три слоя двумерных матриц: оси абсцисс и ординат определяют положение точки (*пикселя*), а три значения по оси аппликат — цвет пикселя. Изображения OCTA B-скан имеют ограниченный цветовой диапазон (рис. 3).

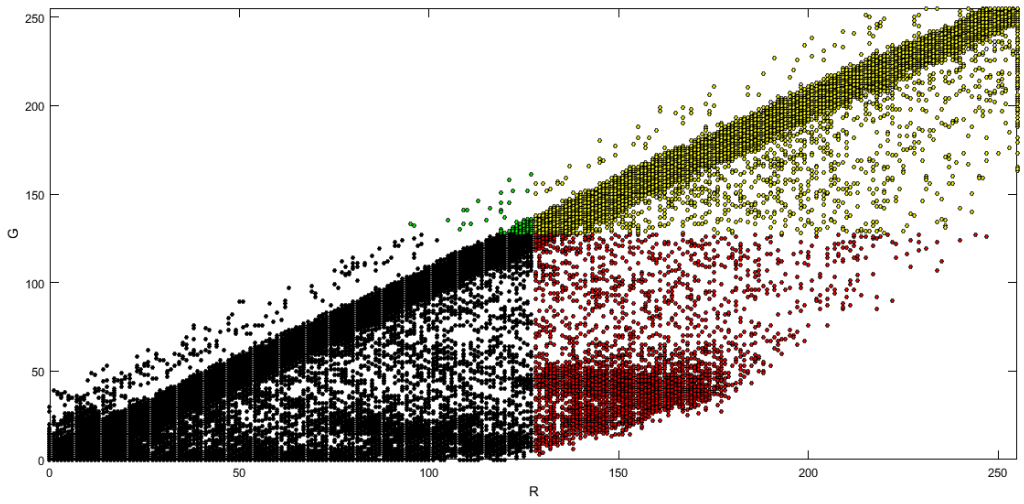

Рис. 3: Исследование цветового диапазона типового OCTA B-скан. По оси абсцисс — значение R в цветовой модели RGB, по оси ординат — значение G в цветовой модели RGB.

Для количественной оценки кровотока в различных слоях сетчатки из исходного изображения в результате обработки выделялось три признака: количество жёлтых пикселей выше OPL слоя, ниже BM и между этими слоями. OPL и BM границы представлены на сканах пунктирными линиями красного цвета. Признаки обозначены как Upper, Under и Middle, соответственно. Можно выделить следующие этапы обработки одного OCTA B-скана:

1. *Выделение границ.* На данном этапе определялось положение (координаты) пикселей красного цвета, соответствующих пунктирным линиям на изображении (Приложение, рис. 1).

2. *Идентификация нижней и верхней границ.* Среди пикселей, выделенных на предыдущем шаге, выбирались те писксели, значения абсцисс и ординат которых, позволяют однозначно отнести их к верхней (OPL) или нижней (BM) границе. Шаг проиллюстрирован на (Приложение, рис. 2).

3. *Дополнение границ.* Полученные значения интерполировались кубическими сплайнами. (Приложение, рис. 3).

4. *Выделение признаков.* На последнем этапе производился подсчет пикселей в интересующих областях. Для наглядности пиксели в разных

областях окрашены в разные цвета: выше OPL границы — в зеленый, между верхней и нижней — в розовый, ниже границы BM — в оранжевый (Приложение, рис. 4). В результате для каждого изображения получены три признака: количество пикселей выше OPL (Upper), ниже BM (Under) и между этими линиями (Middle).

Рисунок 4 позволяет сравнить исходное и обработанное изображения. Для обработки изображений использовался прикладной пакет MATLAB (код в Приложении).

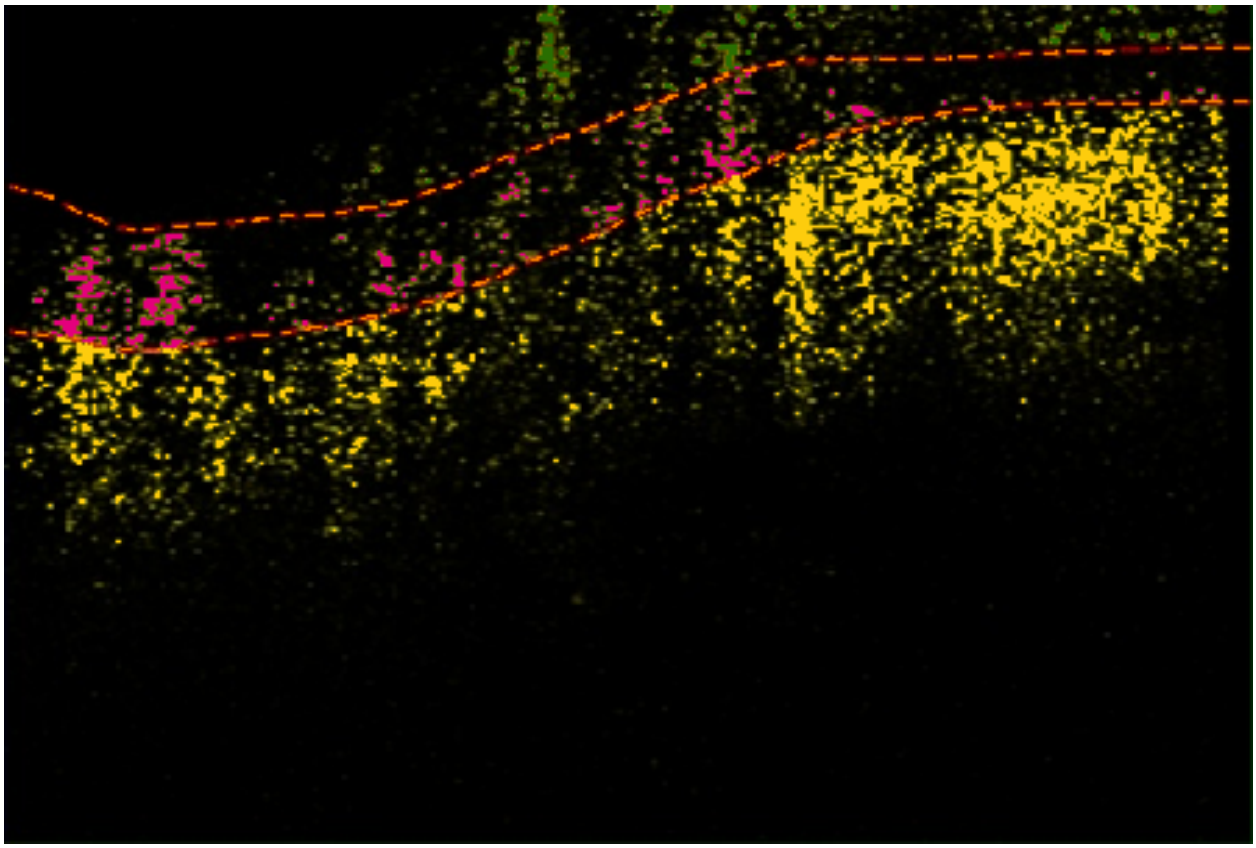

Рис. 4: Сравнение исходного изображения и обработанного путём наложения их друг на друга.

Анализ данных

Структура данных

Всего для анализа были представлены данные 23 пациентов (34 глаза) без офтальмопатологии и с диагнозом "первичная открытоугольная глаукома"(ПОУГ) различных стадий, содержащие 13 признаков. Данные были разбиты на 3 класса с учетом поставленного диагноза: ”Здоров” (23.53%), ”ПОУГ I” (44.12%), ”ПОУГ II или ПОУГ III” (32.35%). Структура данных и доля пропусков представлены в табл. 2.

Подходы к решению

№ п/п	Признак	Значение (пример)	Доля пропусков
1	Middle	4381	нет
2	Upper	9827	нет
3	Under	19280	нет
4	ЦТР	599	17.6%
5	Efrid	0.0343	26.5%
6	Po	19.8	17.6%
7	Pg	18	17.6 %
8	ФРР	9.9	17.6%
9	КГ	8.9	17.6%
10	ТРП	189	35.3%
11	ГРП	406	29.4%
12	SVL	51.47	38.2 %
13	DVL	47.28	47.28%
	Диагноз	ОУГ II	нет

Таблица 2: Структура данных и доля пропусков.

1. Заполнение пропусков осуществлялось выборочным средним по классу, к которому принадлежит объект.

2. Применим фильтрующий метод по критерию Пирсона. Для этого посмотрим на корреляции между признаками на рисунке (Приложение, рис. 5). Как можно увидеть, наблюдается сильная корреляция между 6 (Po) и 7 (Pg) и 7 (Pg) и 8 (ФРР) признаками. Таким образом, следует исключить признаки 6 (Po) и 8 (ФРР).

3. Применим выбор признаков (PCA, LDA, RICA, Lasso) и классификаторы (SVM, Tree, Bagged Tree и AdaBoost), далее сравним результаты. Классификаторы будем сравнивать по метрике MSE по кроссвалидации на 4 фолда.

    - PCA. Посмотрим на объясняемую главными компонентами дисперсию на рисунке (Приложение, рис. 6). Очевидно, что можно оставить только три главные компоненты. Результаты по классификаторам представлены в таблице (Приложение, таблица 1).
    - Lasso. Предварительно ответы на объектах были преобразованы следующим образом: "Здоров- 1, "ОУГ I "ОУГ II + ОУГ III- 0. В результате осталось 8 признаков. Визуализация представлена на рисунке (Приложение,рис. 7). Результаты по классификаторам представлены в таблице (Приложение, таблица 2).
    - LDA. Результаты по классификаторам представлены в таблице (Приложение, таблица 3).

	Tree	SVM	Bagged Tree	Ada Boost
PCA	47.1%	55.9%	55.9%	44.1%
Lasso	61.8%	67.6%	70.6%	55.9%
LDA	91.2%	91.4%	85.3%	44.1%
RICA	50%	64.7%	58.8%	58.8%

Таблица 3: Оценка классификаторов по работе с данными. Сводная таблица.

- RICA. Результаты по классификаторам представлены в таблице (Приложение, таблица 4).

Общие результаты приведены в таблице 3.

4. Проанализируем полученные на предыдущем этапе результаты. Лучшим подходом оказался обработка данных LDA с последующим применением SVM. Посмотрим на данные LDA на рисунке (Приложение, рис. 8). Применение классификатора к другим значениям компонент LDA даёт следующее разделение пространства на три класса (см. рис. 5).

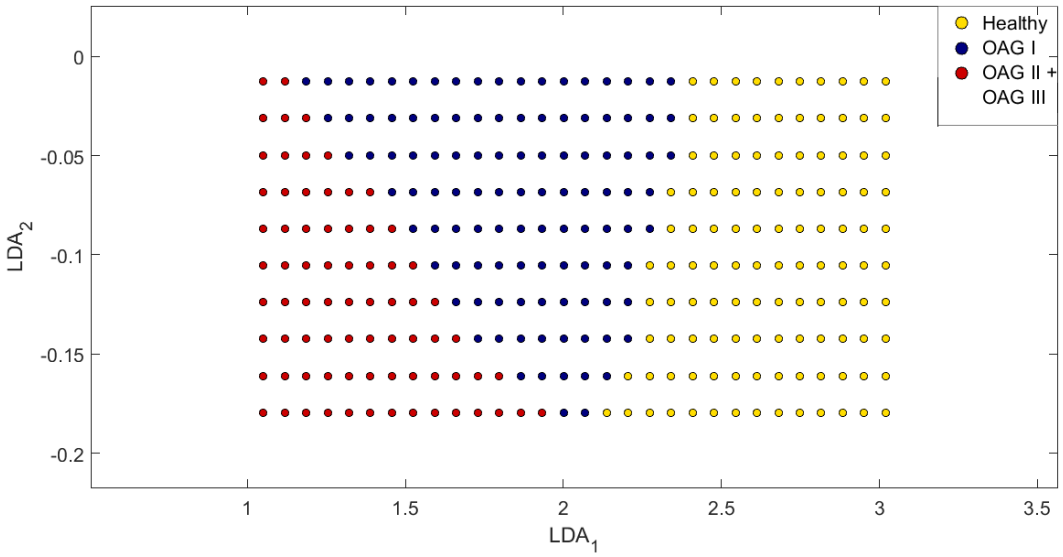

Рис. 5: SVM, применённый к расширенному пространству. Оси абсцисс и ординат обозначают первую и вторую компоненту LDA, соответственно. Цвет означает класс, к которому принадлежит объект.

Выводы

- Был создан алгоритм, позволяющий обрабатывать изображения типа OCTAngio.
- Был подготовлен пайплайн — фильтрация по критерию Пирсона, LDA, SVM — готовый к тому, чтобы оказывать поддержку в клинической практике.

Список литературы

[1] Иомдина Е. Н., Бауэр С. М., Котляр К. Е. Биомеханика глаза: теоретические аспекты и клинические приложения // Под редакцией В.В. Нероева, М.: Реал Тайм, 2015.

[2] Alnawaiseh M. et al Correlation of flow density, as measured using optical coherence tomography angiography, with structural and functional parameters in glaucoma patients // Graefe's Archive for Clinical and Experimental Ophthalmology, Vol.256, No.3, 2018.

[3] Курышева Н. И., Маслова Е. В. Оптическая когерентная томография с функцией ангиографии в диагностике глаукомы // Вестник офтальмологии, Vol.132, No.5, 2016.

[4] Bojikian K. D. et al Optic disc perfusion in primary open angle and normal tension glaucoma eyes using optical coherence tomography-based microangiography // PLoS ONE, Vol.11, No.5, 2016.

[5] Jia Y. et al Optical coherence tomography angiography of optic disc perfusion in glaucoma // Ophthalmology, Vol.121, No.7, 2014.

[6] Geyman L. S. et al Peripapillary perfused capillary density in primary open-angle glaucoma across disease stage: an optical coherence tomography angiography study // British Journal of Ophthalmology, Vol.101, 2017.

[7] Yarmohammadi A. et al Optical coherence comography angiography vessel density in healthy, glaucoma suspect, and glaucoma eyes // Investigative Ophthalmology & Visual Science, Vol.57, No.9, 2016.

[8] Kashani A. H. et al Optical coherence tomography angiography: A comprehensive review of current methods and clinical applications // Progress in Retinal and Eye Research, Vol.60, 2017.

[9] https://www.heidelbergengineering.com/int/press-releases/spectralis-oct-angiography-module-to-be-presented-at-aao/

[10] Ang M. et al Optical coherence tomography angiography: a review of current and future clinical applications // Graefe's Archive for Clinical and Experimental Ophthalmology, Vol.256, No.2, 2018.

[11] Battaglia P. M. et al Vessel density analysis in patients with retinitis pigmentosa by means of optical coherence tomography angiography // British Journal of Ophthalmology, Vol.101, 2017.

[12] Kim A. Y., et al Quantifying microvascular density and morphology in diabetic retinopathy using spectral-domain optical coherence tomography angiography // Investigative Ophthalmology & Visual Science, Vol.57, No.9, 2016.

[13] Hagag A. M., Gao S. S., Jia Y., Huang D. Optical coherence tomography angiography: Technical principles and clinical applications in ophthalmology. // Taiwan Journal of Ophthalmology, Vol.7, No.3, 2017.

[14] Hossin M., Sulaiman M.N. A review on evaluation metrics for data classification evaluations // International Journal of Data Mining & Knowledge Management Process, Vol.5, No.2, 2015.

[15] http://scikit-learn.org/stable/modules/model_evaluation.html#from-binary-to-multiclass-and-multilabel

[16] Cai J., Luo J., Wang S., Yang S. Feature selection in machine learning: a new perspective // Neurocomputing, Vol.300, No.26, 2018.

[17] Chandrashekar G., Sahin F. A survey on feature selection methods // Computers and Electrical Engineering, Vol.40, No.1, 2014.

[18] Rosario S., Thangadurai K. RELIEF: Feature Selection Approach // International Journal Of Innovative Research And Development, Vol.4, No.11, 2015.

[19] Sachin D. Dimensionality reduction and classification through PCA and LDA // International Journal of Computer Applications, Vol.122, No.17, 2015.

[20] Le Q. V. et al ICA with reconstruction cost for efficient overcomplete feature learning // Advances in Neural Information Processing Systems, Vol. 24, 2011.

[21] Tibshirani R. Regression shrinkage and Selection via the Lasso. //Journal of the Royal Statistical Society. Series B (Metodological), Vol.32, No.1, 1996.

[22] Awad M., Khanna R. Support vector machines for classification. // Efficient Learning Machines. Apress, Berkeley, CA. 2015.

[23] http://scikit-learn.org/stable/modules/tree.html#classification

[24] Fratello M., Tagliaferri R. Decision trees and random forests // Reference Module in Life Sciences, Elsevier, 2018.

[25] Heo J., Yang J. Y. AdaBoost based bankruptcy forecasting of Korean construction companies // Applied Soft Computing, Vol. 24, 2014.

[26] Kaushik S., Pandav S. S. Ocular Response Analyzer // Journal of Current Glaucoma Practice, Vol.6, No.1, 2012.

[27] Нестеров А.П., Бунин А.Я., Кацнельсон Л.А. Внутриглазное давление. Физиология и патология. Москва: Наука, 1974.

Приложение

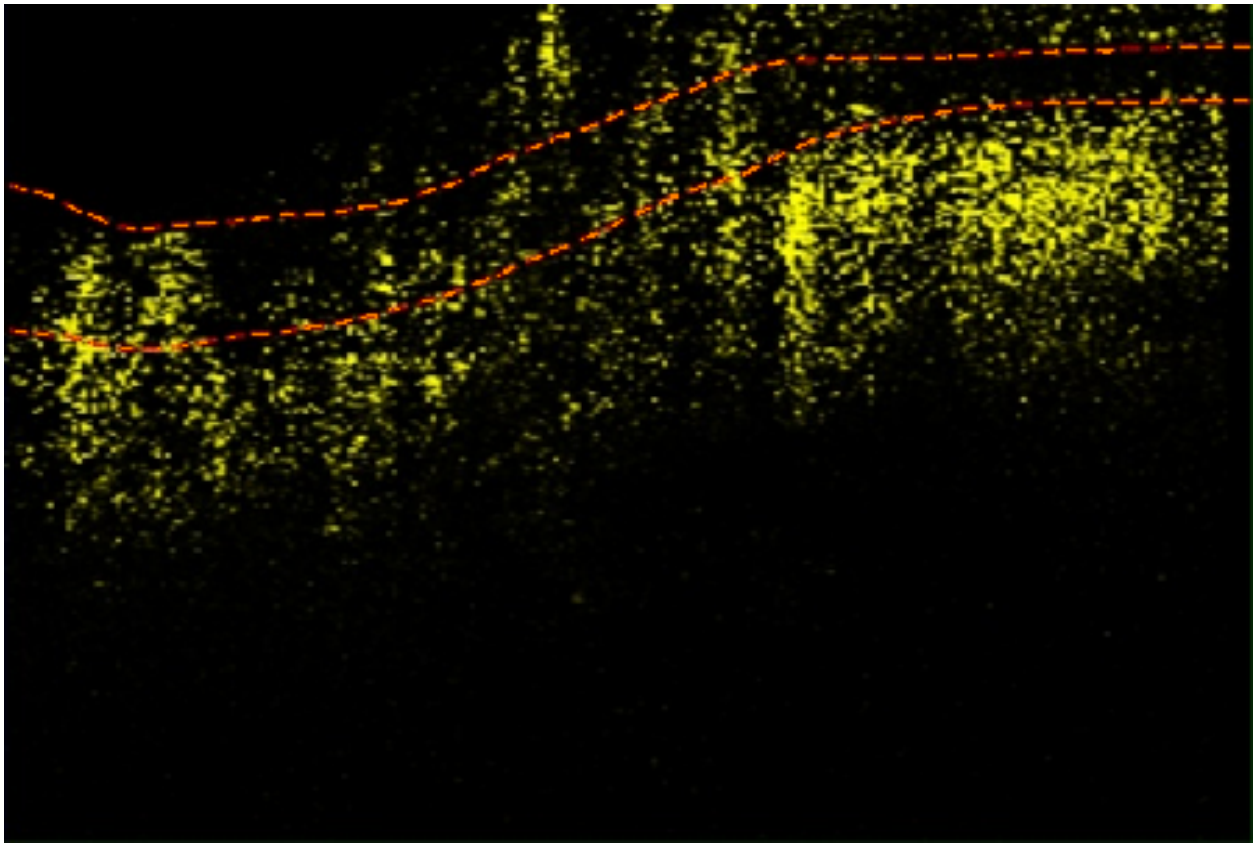

Рис. 1: Обработка OCT скана, шаг 1: Выделение границ.

	Tree	SVM	Bagged Tree	Ada Boost
PCA	47.1%	55.9%	55.9%	44.1%

Таблица 1: Оценка классификаторов по работе с данными после PCA.

	Tree	SVM	Bagged Tree	Ada Boost
Lasso	61.8%	67.6%	70.6%	55.9%

Таблица 2: Оценка классификаторов по работе с данными после Lasso.

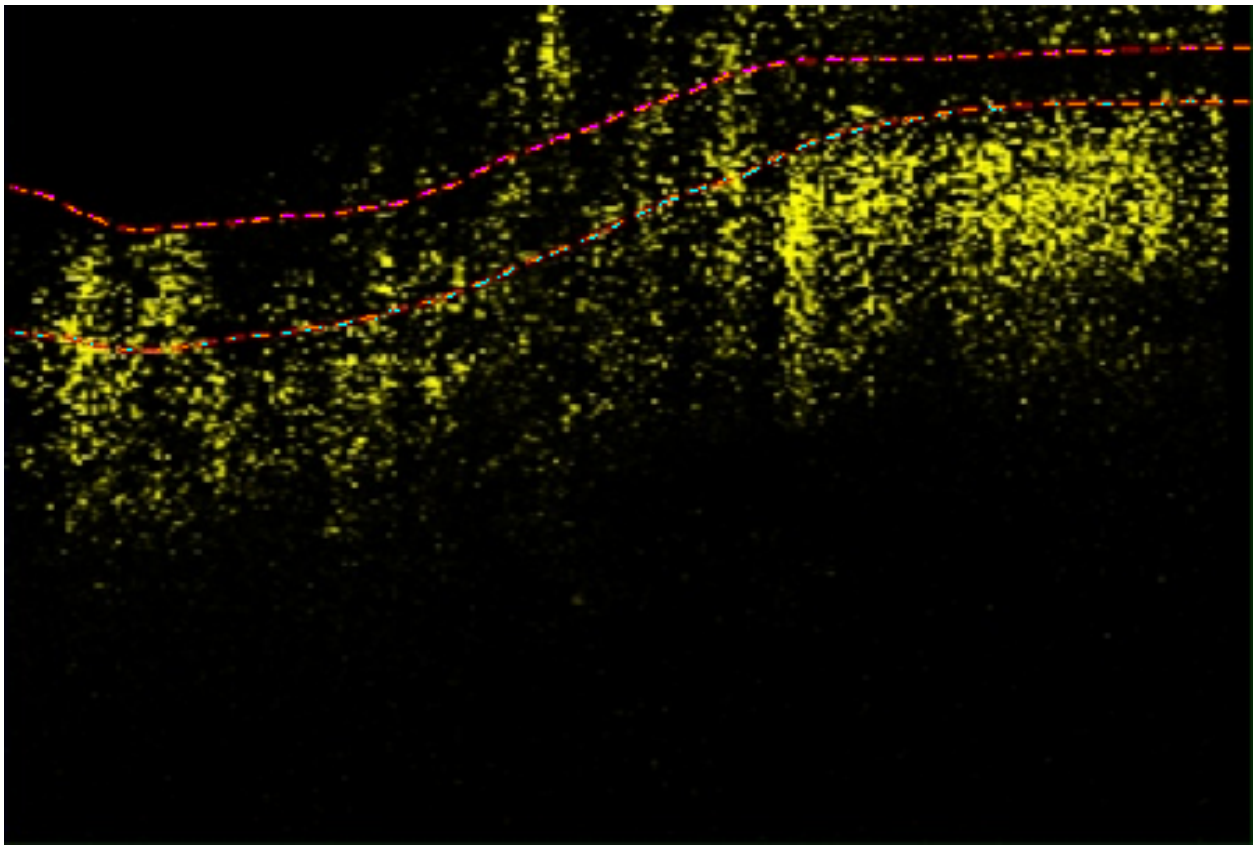

Рис. 2: Обработка OCT скана, шаг 2: Идентификация нижней и верхней границ.

	Tree	SVM	Bagged Tree	Ada Boost
LDA	91.2%	91.4%	85.3%	44.1%

Таблица 3: Оценка классификаторов по работе с данными после LDA.

	Tree	SVM	Bagged Tree	Ada Boost
RICA	50%	64.7%	58.8%	58.8%

Таблица 4: Оценка классификаторов по работе с данными после RICA.

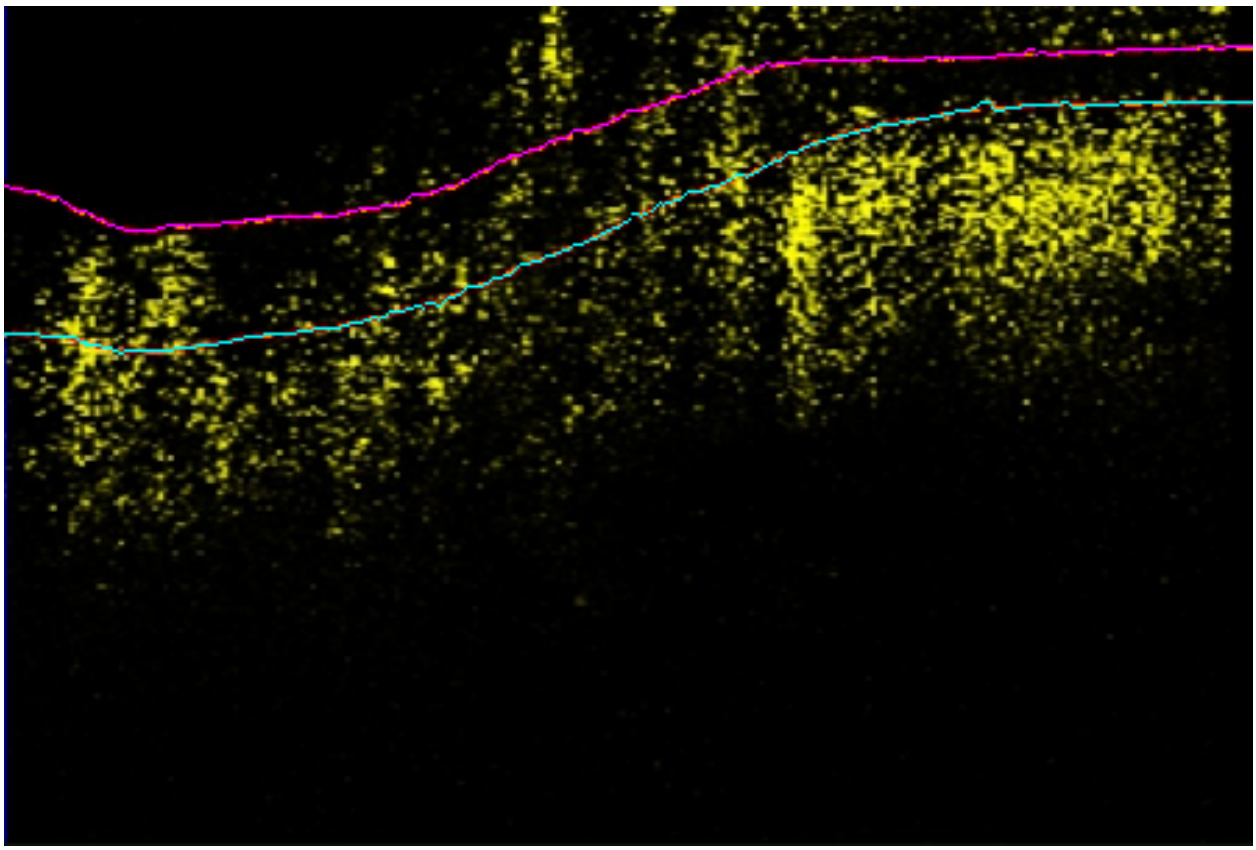

Рис. 3: Обработка ОСТ скана, шаг 3: Дополнение границ.

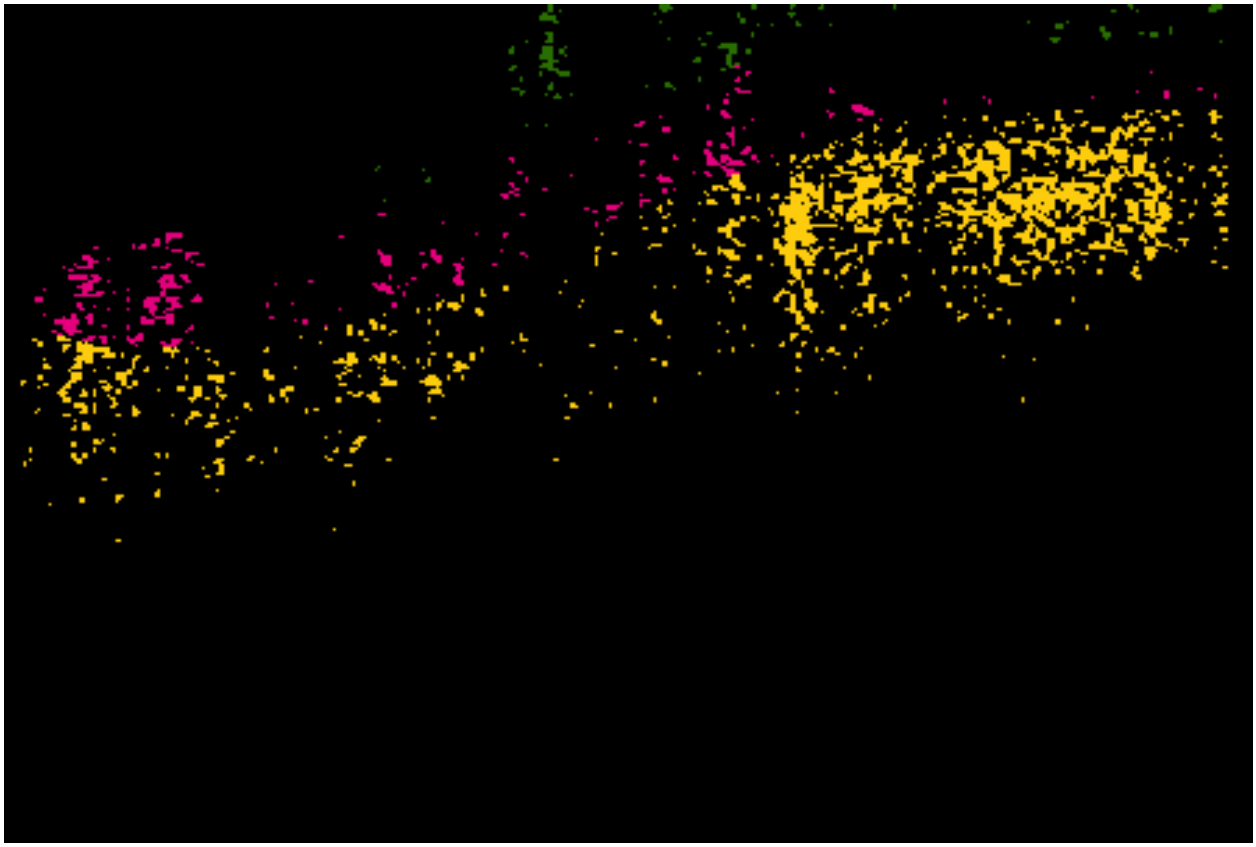

Рис. 4: Обработка ОСТ скана, шаг 4: Выделение признаков.

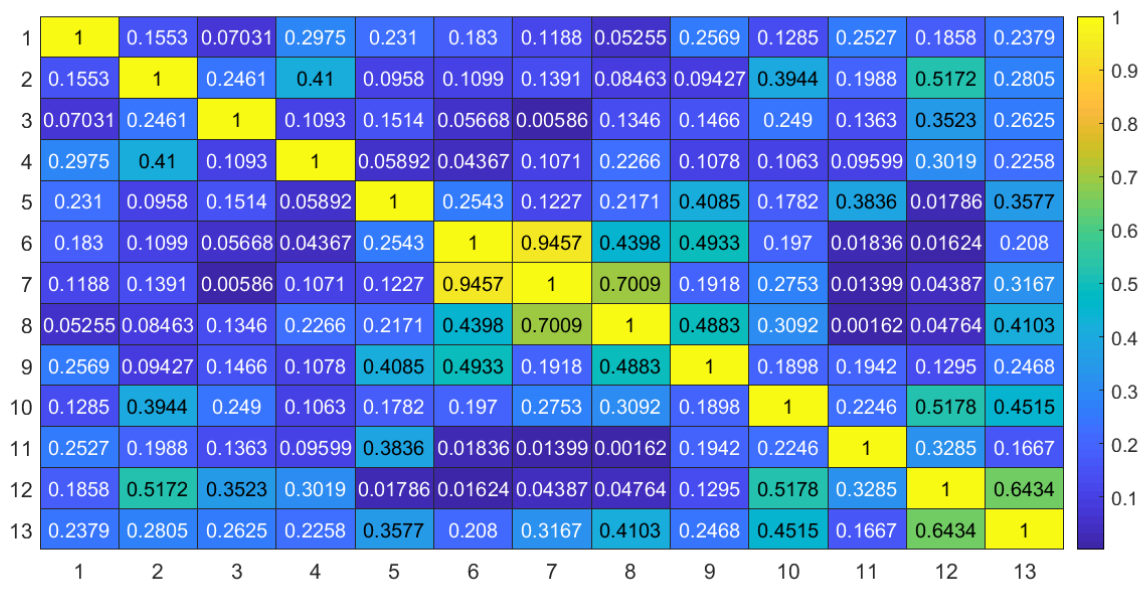

Рис. 5: Абсолютные значения коэффициентов корреляции.

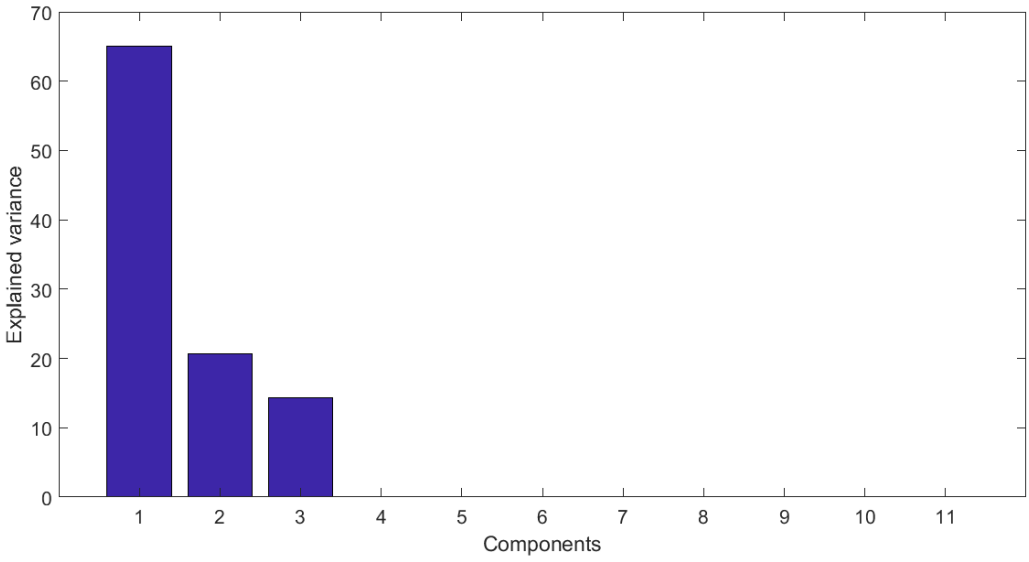

Рис. 6: PCA: дисперсия, объяснённая главными компонентами.

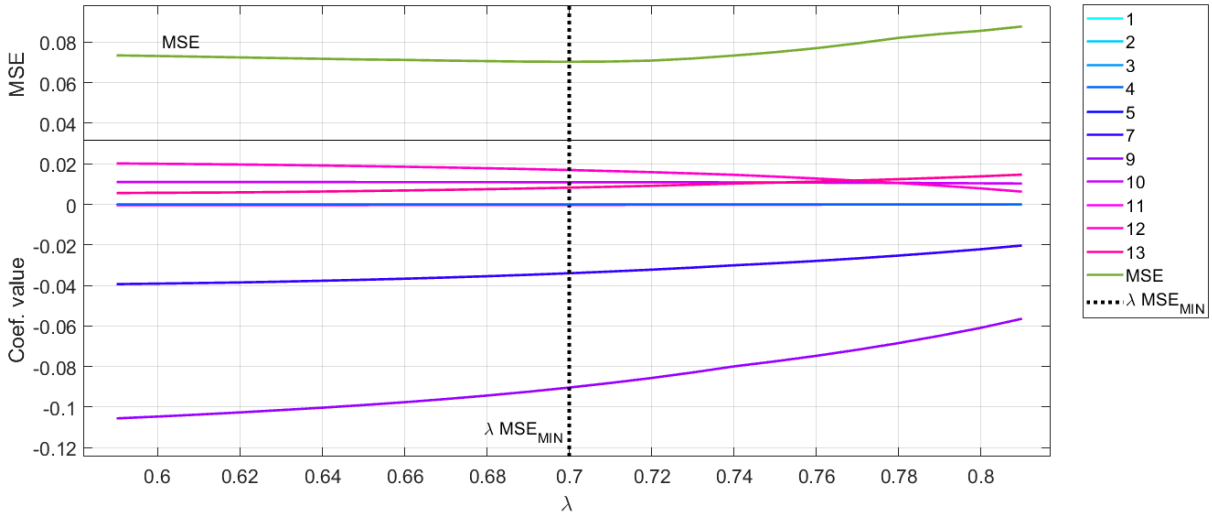

Рис. 7: Lasso: значения коэффициентов и MSE. Чёрной вертикальной пунктирной линией отмечено λ, при котором достигается наименьшее значение MSE.

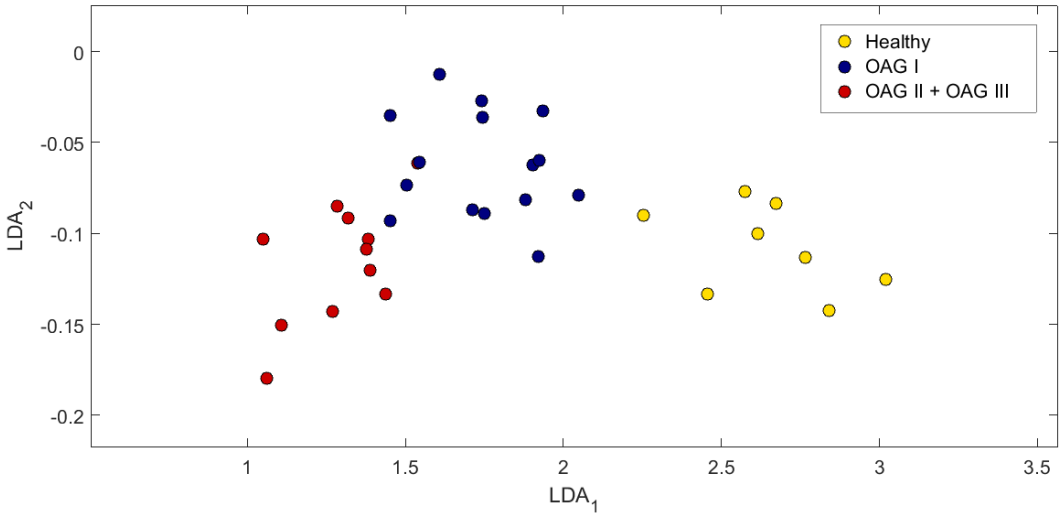

Рис. 8: LDA. Оси абсцисс и ординат обозначают первую и вторую компоненту LDA, соответственно. Цвет означает класс, к которому принадлежит объект.

```
function [Middle,Upper,Under,I1] = nut_pixels(fname)
%% Reading image file
I = imread(fname); H = size(I,1); W= size(I,2);
%% Looking for red border lines
xx=[];
for i=1:H
    for j=1:W
        if ((I(i,j,1)>130)&&(I(i,j,2)<50)&&(I(i,j,1)<200))
            xx=[xx;[i,j]];
        end
    end
end
%% Interpolation of borders
% -- looking for identical W(width) coordinates
% -- and max (min) H(height) coordinates
tab=tabulate(xx(:,2)); ux = tab(tab(:,2)>2,1);
y1 = zeros(size(ux)); y2 = zeros(size(ux));
for i=1:length(ux)
    y1(i) = min( xx(xx(:,2)==ux(i),1) );
    y2(i) = max( xx(xx(:,2)==ux(i),1) );
end
% -- cubic spline interpolation
yy1 = floor(spline(ux,y1,[1:W])); yy2 = floor(spline(ux,y2,[1:W]));
%% Pixel count and returning processed image
Upper=0; Middle=0; Under=0; Middle_black=0; I1=I;
for j=1:W
    for i=1:H
        if((I1(i,j,1)>127)&&(I1(i,j,2)>127)&&(I(i,j,3)<125)) % if yellow
            if (i<yy1(j)) % upper
                Upper = Upper+1;
                I1(i,j,1)=41 ; I1(i,j,2)=110 ; I1(i,j,3)=1;
            elseif (i>yy2(j)) % under
                Under = Under+1;
                I1(i,j,1)=253 ; I1(i,j,2)=204 ; I1(i,j,3)=13;
            else % middle
                Middle = Middle+1;
                I1(i,j,1)=228 ; I1(i,j,2)=0 ; I1(i,j,3)=124;
            end
        else % if not yellow
            I1(i,j,1)=0 ; I1(i,j,2)=0 ; I1(i,j,3)=0;
        end
    end
end
end
```